Clinical sequencing: From raw data to diagnosis with lifetime value

Caspar, S M ; Dubacher, N ; Kopps, A M ; Meienberg, J ; Henggeler, C ; Matyas, G

Abstract: High-throughput sequencing (HTS) has revolutionized genetics by enabling the detection of sequence variants at hitherto unprecedented large scale. Despite these advances, however, there are still remaining challenges in the complete coverage of targeted regions (genes, exome or genome) as well as in HTS data analysis and interpretation. Moreover, it is easy to get overwhelmed by the plethora of available methods and tools for HTS. Here, we review the step-by-step process from the generation of sequence data to molecular diagnosis of Mendelian diseases. Highlighting advantages and limitations, this review addresses the current state of (1) HTS technologies, considering targeted, whole-exome, and whole-genome sequencing on short- and long-read platforms; (2) read alignment, variant calling and interpretation; as well as (3) regulatory issues related to genetic counseling, reimbursement, and data storage.

DOI: https://doi.org/10.1111/cge.13190

Received: 6 October 2017 | Revised: 28 November 2017 | Accepted: 30 November 2017

DOI: 10.1111/cge.13190

WILEY CLINICAL GENETICS

INVITED REVIEW

Clinical sequencing: From raw data to diagnosis with lifetime value

S.M. Caspar[1] | N. Dubacher[1] | A.M. Kopps[1] | J. Meienberg[1] | C. Henggeler[1] | G. Matyas[1,2]

[1]Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Schlieren-Zurich, Switzerland

[2]Zurich Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland

Correspondence
Dr Gabor Matyas, Center for Cardiovascular Genetics and Gene Diagnostics, Foundation for People with Rare Diseases, Wagistrasse 25, CH-8952 Schlieren-Zurich, Switzerland.
Email: matyas@genetikzentrum.ch

Funding information
COFRA Foundation; Ernst Göhner Stiftung; Foundation Suyana; Gebauer Stiftung; Gottfried & Julia Bangerter-Rhyner-Stiftung; NOMIS Stiftung; Palatin-Stiftung; Schäppi-Jecklin Stiftung

High-throughput sequencing (HTS) has revolutionized genetics by enabling the detection of sequence variants at hitherto unprecedented large scale. Despite these advances, however, there are still remaining challenges in the complete coverage of targeted regions (genes, exome or genome) as well as in HTS data analysis and interpretation. Moreover, it is easy to get overwhelmed by the plethora of available methods and tools for HTS. Here, we review the step-by-step process from the generation of sequence data to molecular diagnosis of Mendelian diseases. Highlighting advantages and limitations, this review addresses the current state of (1) HTS technologies, considering targeted, whole-exome, and whole-genome sequencing on short- and long-read platforms; (2) read alignment, variant calling and interpretation; as well as (3) regulatory issues related to genetic counseling, reimbursement, and data storage.

KEYWORDS

genetic counseling, genetic testing, long-read sequencing, next-generation sequencing, pharmacogenetics, short-read sequencing, targeted gene panels, WES, WGS

1 | INTRODUCTION

Prior to the current genomics era, exon-by-exon Sanger sequencing[1] has been used for the sequence analysis of a single or a few genes. In the last decade, while Sanger sequencing has remained the gold standard for confirming sequence variants, high-throughput sequencing (HTS), also known as next-generation sequencing (NGS) has revolutionized genetics by massive parallelization of sequencing reactions, leading to a throughput several orders of magnitude higher than Sanger sequencing.[2] This high throughput allows to sequence a specific panel of genes (targeted sequencing, TS), the entire genome (whole-genome sequencing, WGS) or its coding part (whole-exome sequencing, WES) in a matter of hours to days, depending on the technology and protocol used. The applied HTS technologies and analysis pipelines, however, determine not only the time frame but also the sensitivity/recall, precision, and disk footprint of variant calling as well as the type of detectable sequence variants. It is therefore of particular importance to be aware of the available HTS methods and analysis tools for a best practice workflow in clinical sequencing.

Here, according to the recent literature and our own data, we review the workflow from sequence data generation to molecular diagnosis of Mendelian diseases (Figure 1), showing the advantages and limitations of the most widely used HTS technologies and analysis tools. Our review is divided into 3 main sections: (1) HTS technologies, (2) data analysis and interpretation, and (3) regulatory issues. Accordingly, the first section addresses the current state and future trends of HTS technologies, considering TS, WES, and WGS on short- and long-read platforms. The second section overviews read alignment (ie, mapping of reads to the reference genome) as well as the ability to call different types of genomic sequence variants, such as single-nucleotide variants (SNVs), small insertions and deletions (indels, ≤50 bp), copy number variations (CNVs, >50 bp), and short tandem repeats (STRs). The filtering and interpretation of called sequence variants are addressed in this section as well. We focus on a selection of software tools, being aware that laboratories may use their own in-house data analysis and interpretation pipeline. The third section is concerned with regulatory issues related to genetic

Sylvan M. Caspar and Nicolo Dubacher contributed equally to this work.

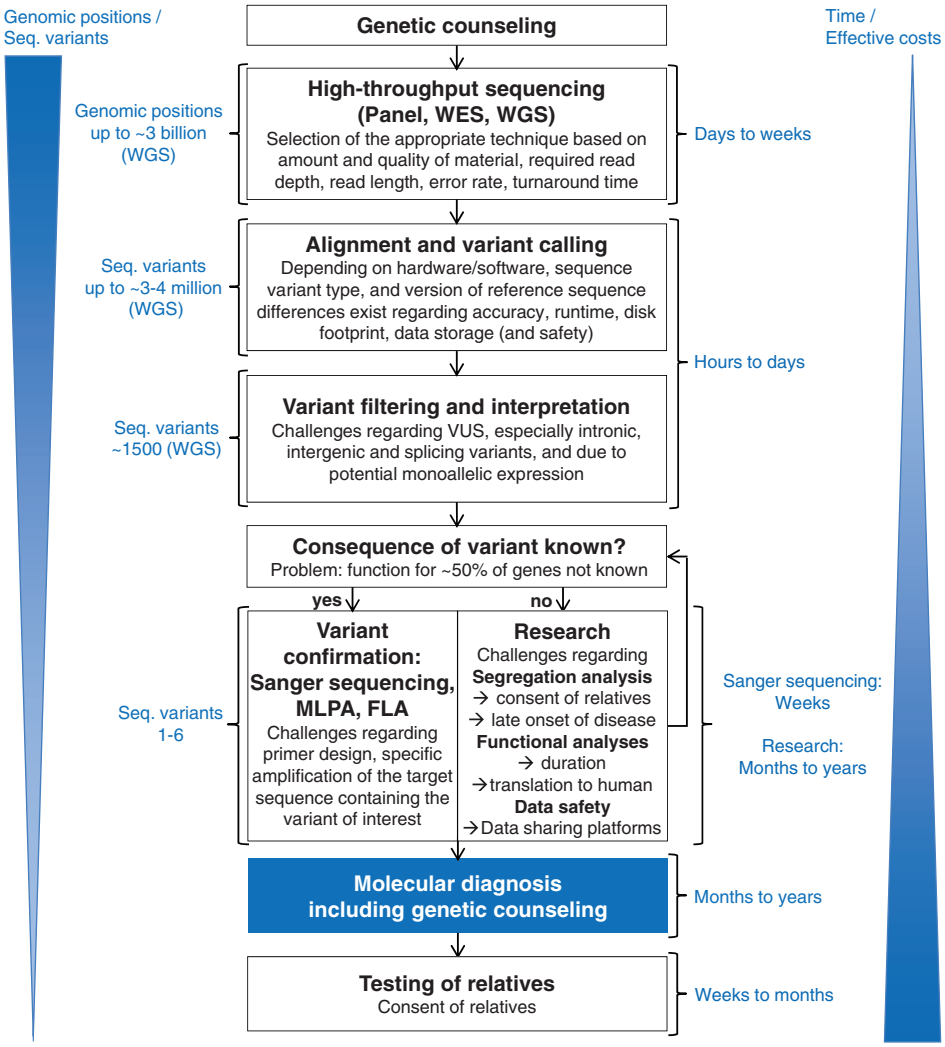

FIGURE 1 Flow chart summarizing the process from sequencing to diagnosis, including associated challenges during each step. FLA, fragment length analysis; MLPA, multiplex ligation-dependent probe amplification; Seq. variants, sequence variants; VUS, variants of unknown significance; WES, whole-exome sequencing; WGS, whole-genome sequencing

counseling, reimbursement, and data storage. A glossary including abbreviations is provided (Table 1).

2 | HIGH-THROUGHPUT SEQUENCING TECHNOLOGIES

2.1 | Targeted sequencing

Genes of interest, so-called gene panels, can be enriched and sequenced simultaneously by TS, which serves as an inexpensive and rapid first-tier test with compatibility to benchtop short-read sequencers (eg, Illumina MiSeq/NextSeq and Thermo Fisher Scientific Ion PGM/Proton) and with especially high read depth of targeted regions (Table 2). Consequently, TS enables sequence variant detection in samples with very low non-reference allele frequencies caused by germline mosaics or clonal complexity of tumors.[3,4] Moreover, the use of disease-related gene panels facilitates the interpretation of detected sequence variants and minimizes the chance of incidental findings.[5] According to the guidelines and recommendations of the American College of Medical Genetics and Genomics (ACMG), only genes with a scientifically sufficiently supported role in disease should be included in a clinical gene panel.[6]

Limitations of TS include the difficulty of detecting clinically relevant sequence variants, especially CNVs, in regions with an insufficient number of reads (ie, below the limit of variant calling) resulting in discontinuous coverage (<100%) (Table 2). Incomplete coverage can result from poor enrichment of GC-rich regions such as first exons and from the absence of probes used for enrichment (Figure 2).[7] Furthermore, due to the identification of novel gene-disease associations, gene panels require updates, especially in diseases with incompletely understood molecular basis. Thus, TS may be inconclusive in the case of negative results, providing no diagnosis and requiring a different gene panel or an additional method for CNV detection, or second-tier sequencing (WES and/or WGS). Updating gene panels can be avoided by sequencing all (known) genes (~25 000) applying WES or WGS.

2.2 | Whole-exome sequencing

As the exome encompasses ~2% of the human genome but harbors ~85% of all described disease-causing sequence variants,[8,9] WES aims to sequence the coding exons of all our known genes, that is, considerably more than TS. The so-called clinical exome encompasses ~5000 disease-associated genes (eg, Agilent SureSelect Focused Exome, Illumina TruSight One and Roche NimbleGen SeqCap EZ

TABLE 1 Glossary and abbreviations

BAM (binary alignment map): file format for storing sequence reads aligned to the reference genome (compressed binary representation of SAM).
Benchmark reference materials: samples with known and validated sequence variants, which can be used for the benchmarking and evaluation of sequencing platforms and software tools.
Bioinformatic pipelines: workflow combining a variety of software tools for biological data analysis.
Clinical exome: HTS restricting sequencing to disease-associated genes.
ClinVar: a freely accessible, public archive for the clinical interpretation of human sequence variants (ncbi.nlm.nih.gov/clinvar).
CNV: copy number variation (deletions and duplications >50 bp).
CRISPR-Cas9-targeted enrichment: a novel, amplification-free enrichment technique that employs the CRISPR-Cas9 system for specific targeting of multiple genomic loci.[105]
De novo assembly for CNV detection: approach to identify SVs by merging and ordering unaligned reads to reassemble the original sequence from which the reads were sampled.
Disk footprint: storage footprint on a disk expressed in bytes such as mega- (MB), giga- (GB) or terabytes (TB).
Enrichment bias: in HTS with a hybridization-based capture and/or amplification step, certain genomic regions are enriched with different efficiency (eg, due to differences in GC content). Indeed, GC-rich regions are difficult to enrich, leading to a below-average number of reads and to incomplete coverage of targeted regions (genes, exome or genome).
Enrichment kits: reagent kits for capturing pre-selected (targeted) genomic regions of interest (eg, exome).
ExAC/gnomAD: the Exome Aggregation Consortium (ExAC; exac.broadinstitute.org) and the Genome Aggregation Database (gnomAD; gnomad.broadinstitute.org) are large-scale sequencing projects currently containing 60 706 exomes and 123 136 exomes/15 496 genomes, respectively. These data sets serve as the reference for population frequencies but may include individuals with late-onset or unrecognized disease.
FASTQ format: text-based format for storing both biological sequences (usually nucleotide sequences) and their corresponding quality scores.
FN (false-negative sequence variant): variant missed by variant calling, despite being present. The goal of clinical sequencing should be to have zero FNs as they could lead to missed diagnoses.
FP (false-positive sequence variant): variant detected by variant calling, despite being absent. The number of FPs can be reduced by appropriate (eg, frequency-based) variant filtering.
GC-rich regions: genomic regions containing high numbers of guanine (G) and cytosine (C). Due to the base stacking, GC-rich genomic regions are particularly stable and hence more difficult to enrich.
HGMD: the commercial Human Gene Mutation Database constitutes a comprehensive core collection of data on germline sequence variants in nuclear genes underlying or associated with human inherited disease (portal.biobase-international.com). The less up-to-date public version of HGMD is freely available for academic institutions/non-profit organizations (www.hgmd.cf.ac.uk).
HPO: the Human Phenotype Ontology is a database aiming to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease (human-phenotype-ontology.github.io).
HTS: high-throughput (formerly next-generation) sequencing.
Incidental findings: sequence variants with potential health or reproductive importance but unrelated to the clinical phenotype for which sequencing was performed.
Indels: small insertions and deletions (≤50 bp).
Interpretation databases: Examples for SNVs/indels: OMIM, HGMD, ClinVar, LSDB, ExAC/gnomAD; for CNVs: DGV (dgv.tcag.ca).
Interpretation software: Examples for SNVs/indels: Alissa Interpret (agilent.com/lifesciences/alissa; suitable for CNVs as well), Ingenuity Variant Analysis (qiagenbioinformatics.com/products/ingenuity-variant-analysis), Exomiser/Genomiser (exomiser.github.io/Exomiser), NxClinical (biodiscovery.com/nxclinical; suitable for CNVs as well), VarSeq (goldenhelix.com/products/VarSeq); for CNVs: InHelix (diploid.com), Nexus copy number (BioDiscovery); for manual interpretation: Alamut Visual (interactive-biosoftware.com), Varsome (varsome.com).
LRS: long-read sequencing.
LSDB: locus specific mutation database (hgvs.org/locus-specific-mutation-databases).
Mappability: computed value giving the inverse of the number of times that a read maps to the genome for a given read length and number of allowed mismatches (eg, $m = 2$). A mappability of 1 indicates unambiguous mappable regions, whereas a mappability <1 indicates regions which tend to produce ambiguous mappings.[33]
Non-reference allele: allele with sequence variant, that is, with sequence differing from the reference genome.
OMIM: Online Mendelian Inheritance in Man is a freely available, comprehensive, authoritative compendium of human genes and genetic phenotypes, containing information on all known Mendelian disorders and currently over 15 000 genes (omim.org).
ONT: Oxford Nanopore Technologies (nanoporetech.com).
PacBio: Pacific Biosciences (pacb.com).
Paired-end mapping: approach to identify SVs by evaluating the span and orientation of paired-end reads. CNVs are indicated by read pairs with mapping spans inconsistent with the expected insert size.
Precision: analysis quality measure (TP/[TP + FP]), analogous to positive predictive value.
Random monoallelic expression (MAE): only 1 of 2 alleles of a gene is actively transcribed, while the other allele is silent. MAE can occur in a random fashion, varying in an interindividual-, tissue- and/or cell-type-specific manner[87] and its evaluation is assisted by the database of monoallelic expression (dbMAE, mae.hms.harvard.edu), containing information on tissue-specific expression patterns of autosomal genes.[88]
Read alignment: mapping/assigning of raw short or long reads (sequencer output stored in FASTQ or related formats) to the reference genome. Software examples for the alignment of short reads: BWA-MEM,[63] Isaac,[65] GENALICE MAP (genalice.com).[60]
Read depth: number of aligned sequencing reads covering a specific genomic position.

(*Continues*)

TABLE 1 (Continued)

Read-depth analysis: approach to identify CNVs by comparing the read depth of genomic regions in a sample to its own average read depth or to the read depth of the same region in one or multiple control samples.
Reference genome: the current human reference genome (assembly of a number of donors) is GRCh38 (the Genome Reference Consortium human genome build 38) or hg38 (equivalent version). Reference genomes are improved by the Genome Reference Consortium and these improvements should lead to more accurate genomic analyses.[106] The previous human reference genome GRCh37/hg19 is still widely used.
RefSeq: the Reference Sequence collection (ncbi.nlm.nih.gov/refseq) provides a comprehensive, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies.
SAM (sequence alignment map): file format for storing biological sequences aligned to the reference genome. It is widely used for storing data, such as nucleotide sequences. The format supports short and long reads.
Sensitivity/recall: analysis quality measure (TP/[TP + FN]), analogous to true positive rate or probability of detection.
SNP (single-nucleotide polymorphism): variation in a nucleotide at a single specific position among individuals, occurring with appreciable relative frequency (e.g. >1%) within a population (not to be confused with SNV).
SNV (single-nucleotide variant): variation in a nucleotide at a single specific position between individual and reference genomes without any frequency limitation (not to be confused with SNP). dbscSNV is a database for SNVs within splicing consensus regions.[82]
Split reads: approach to identify SVs by investigating breakpoints in reads split into two parts and aligned separately to the genome (different position and/or orientation).
SRS: short-read sequencing.
SRS dead zone: genomic region for which multiple matches of short-read sequencing reads exist, making unambiguous mapping impossible.[35]
STR (short tandem repeat): DNA sequence with a DNA motif of a few base pairs repeated in tandem (also known as microsatellite).
SV (structural variation): sequence variant affecting the structure of a chromosome (eg, copy number variations, insertions, and translocations).
TP (true-positive sequence variant): sequence variant which are present and detected by variant calling.
Trio analysis: HTS data of an index case and his/her parents is used to facilitate the detection of disease-causing sequence variant(s).
TS (targeted sequencing): a rapid and cost-effective way to detect known and novel variants in selected/enriched sets of genes or genomic regions (specified in gene panels).
Variant calling: process of identifying sequence variants by detecting positions differing between sample sequence and reference genome. Software examples for *SNVs/Indels*: GATK-HC,[64] Isaac,[65] or GENALICE MAP (genalice.com); for CNVs: CNVnator,[69] BreakDancer,[70] LUMPY,[71] Manta,[72] Cortex[73]; for STRs: lobSTR,[77] ExpansionHunter,[78] HipSTR,[79] or STRetch.[80]
(g)VCF: (genomic) Variant Call Format used for storing genomic sequence variants. (g)VCF files contain meta-information lines, a header line, and data lines containing information about the called sequence variants and their positions in the genome.
VUS: sequence variants of unknown significance.
WES: whole-exome sequencing.
WGS: whole-genome sequencing.

MedExome), thereby covering only ~20% of the entire exome and thus requiring additional analyses if the disease-causing sequence variant remains undetected.[10,11] In contrast, WES not only covers exons of already disease-associated genes but also allows for the identification of novel gene-disease associations in diseases with yet unknown molecular basis in a cost-effective way.[12,13] It is therefore not surprising that WES is widely used and has been advocated as a first-tier test instead of TS.[14,15] The cost efficiency of WES makes trio analysis feasible (ie, sequencing the patient and his/her parents), which facilitates data interpretation and increases the diagnostic yield considerably (from ~20% to ~30%).[16,17]

WES data analysis can be limited in silico to a gene panel (genes of interest) and subsequently expanded if necessary, thereby reducing the potential for the discovery of incidental/secondary findings.[18,19] Although WES can detect more sequence variants than TS, WES has similar limitations as TS (Figure 2, Table 2). Accordingly, WES may fail to cover poorly enriched parts of the exome and hence to detect clinically relevant sequence variants, that is, ~400 exonic disease-causing variants listed in HGMD (portal.biobase-international.com).[7] The incomplete coverage of WES may be improved by combining enrichment kits (eg, Agilent SureSelect, Illumina TruSeq Capture, and Roche NimbleGen SeqCap EZ) or by using a high concentration of capture probes that cover difficult-to-enrich regions.[20–23]

2.3 | Whole-genome sequencing

WGS has the advantages of the most continuous coverage and identifying sequence variants throughout the genome. These advantages enable WGS (1) to be a better WES[7]; (2) to detect non-exonic sequence variants[7,21] and (3) to improve CNV detection,[7,24] leading to increased diagnostic yield over WES.[25–28]

WGS data interpretation, as in WES, may initially focus on a region of interest by applying in silico gene panels. To date, deep intronic, intergenic, and regulatory sequence variants are difficult or impossible to interpret at the DNA level only.[29] However, even such sequence variants can be of importance in the near or distant future, adding a superior value to WGS data and putting the higher costs of WGS into a lifetime perspective. This superior value of WGS can be recognized in the emerging field of pharmacogenetics.[30] Pharmacogenetic predisposition, which can be obtained by subsequent filtering of appropriate HTS data, is clinically highly relevant, because it can play a pivotal role in the success or failure of a pharmacological therapy. Notably, some pharmacogenetically-relevant sequence variants, such as a single nucleotide polymorphism (SNP) in the promoter region of *VKORC1* (rs9923231) causing low-warfarin-dose phenotype[31] or structural variations (SVs) affecting drug targets,[32] can readily be detected by WGS but not by WES. The limitation of WGS is mainly

TABLE 2 Comparison of widely used sequencing applications and platforms

		Short-read[a]			Long-read (real) WGS		
	Sanger	TS	WES	PCR-free WGS	PacBio	ONT	References
Read length (bp)	Max: 500-1000	~300	~150	~150	Up to template length[b]	Up to template length[b]	1,2,38,41
Typical read depth	Not applicable	200-1000×	~100×	30-60×	10-30×	10-30×	3,7,20,37,42
Raw-read error rate (%)	0.001	0.1	0.1	0.1	10-15	12-17	2,38,47
Costs per sample ($)[c]	15-20	200-1000	500-1000[f]	1000-2500[f]	7000-20 000[g]	2750-8250[d]	
Disk footprint (GB)/($)[e]	<0.1/0.01	<1/<0.1	6-13/<1	90-400/4-20	45-130/2-7	75-220/4-11	107
Advantage	High accuracy	High read depth, easy interpretation, cost efficiency, short turnaround time	Additional sequence information compared to TS, cost efficiency	Uniform, GC content-independent coverage of the genome	Coverage of repetitive and homologous genomic regions, detection of large SVs, discovery of novel isoforms, DNA/RNA base modifications, phasing		3,4,7,14,20,38,41
Limitation	Low throughput	Incomplete coverage due to high GC content, missing enrichment probes, and regions with mappability<1	Incomplete coverage due to high GC content, missing enrichment probes, and regions with mappability <1	Incomplete coverage in regions with mappability <1	High first-pass (raw-read) error rate, low cost efficiency		7,20,38,41
Amplification step prior to sequencing	Yes	Yes	Yes	No	No	No	1,3,4,7,20

Abbreviations: ONT, Oxford Nanopore Technologies; PacBio, Pacific Biosciences; PCR, polymerase chain reaction; SV, structural variation; TS, targeted sequencing; WES, whole-exome sequencing; WGS, whole-genome sequencing.

[a] Parameters of short-read sequencing are adapted to Illumina MiSeq v3 system (TS) and Illumina HiSeq X Ten system (WES, WGS). Note that reads cannot be unambiguously aligned to repetitive/homologous regions larger than the read length, that is, mappability <1 (cf. Figure 3).

[b] Maximal read length only limited by length of the fragments sequenced (template).

[c] Costs calculated according to most frequently used sequencing systems, library preparation kits, and reagents for the respective application, considering "typical read depth."

[d] According to nanoporetech.com/about-us/news/human-genome-minion.

[e] Calculated for files like FASTQ, BAM, and VCF using corresponding in-house and publicly available data (github.com/nanopore-wgs-consortium/NA12878, ftp-trace.ncbi.nih.gov/giab/ftp/data/AshkenazimTrio). As file sizes can vary substantially according to the sequencing platform, read depth, and analysis software used,[57–60] storage footprint (in GB) is estimated/exemplified for typical cases. Costs were calculated as described elsewhere,[107] considering disk footprint for backup as well. For TS, disk footprint was calculated for 100 average-sized genes with 2.5-kb coding region per gene.

[f] According to illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf.

[g] According to allseq.com/knowledge-bank/sequencing-platforms/pacific-biosciences.

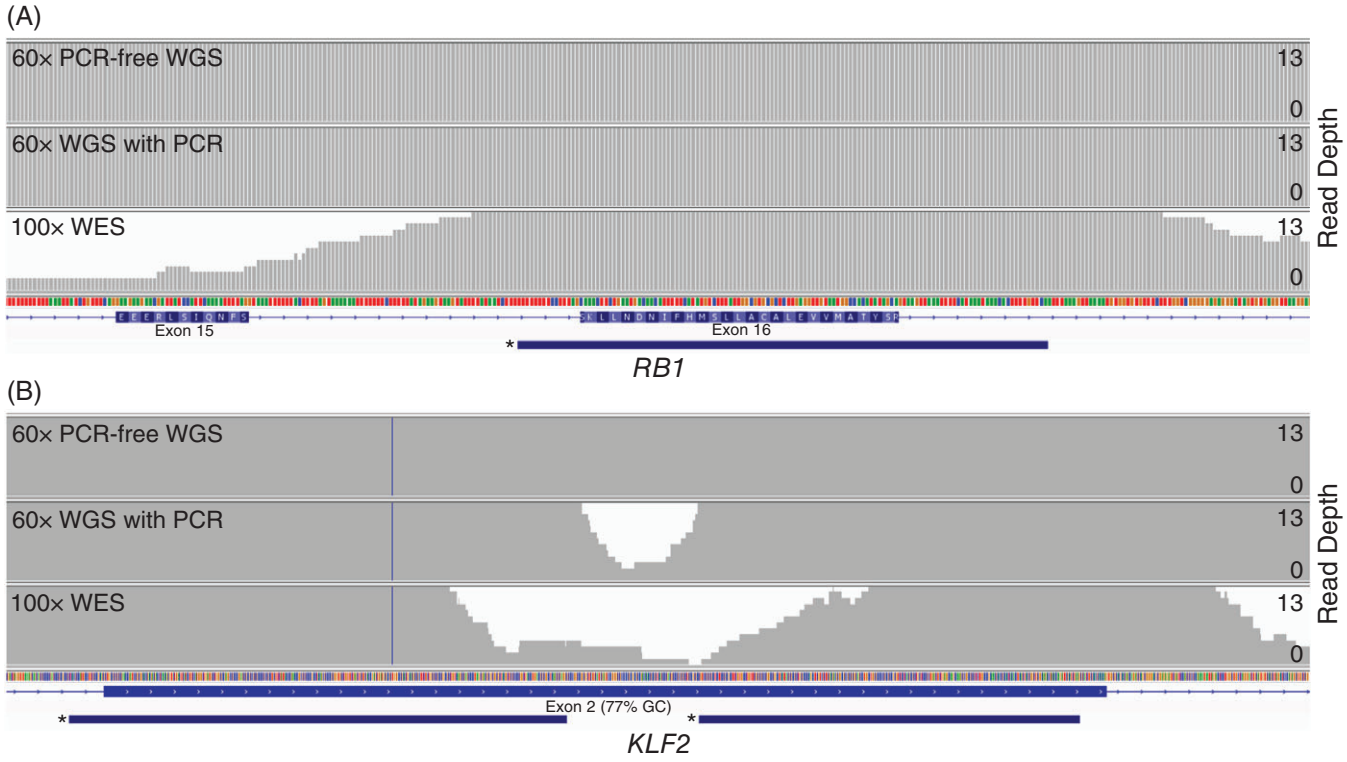

FIGURE 2 Coverage difference between WES and WGS exemplified for 2 genes (adapted and modified from Reference 7). (A) Not all exons are captured leading to low/insufficient read depth in WES in contrast to WGS. (B) GC-rich (77%) exon 2 of *KLF2* is completely covered by PCR-free WGS and almost completely by WGS with PCR, while a large gap in coverage is present in WES. Coverage tracks are visualized by the Integrative Genomics Viewer (IGV, broadinstitute.org/igv). For abbreviations, see Figure 1 and Table 2. *Designed target region of Agilent SureSelect v5 + UTR

determined by the sequencing read length leading to uncertain (<1) mappability (Table 2).

2.4 | Short-read sequencing

TS, WES, and WGS have traditionally been performed using short-read sequencing (SRS), also known as second-generation sequencing. SRS typically produces reads that are 100 to 400 bp in length, depending on the used technology. The current SRS market leader is Illumina (illumina.com) but also other manufacturers offer SRS sequencers, such as Thermo Fisher Scientific (thermofisher.com) and Beijing Genomics Institute (bgi.com) offering Ion Torrent and proprietary BGISEQ (former Complete Genomics) platforms, respectively.

SRS is typically based on library preparation by random fragmentation of input DNA and subsequent adapter ligation, followed by massively parallel sequencing of adapter-ligated fragments.[2] While the main advantages of SRS include high-throughput, low per base cost and low raw-read error rate (Table 2), the common limitation of

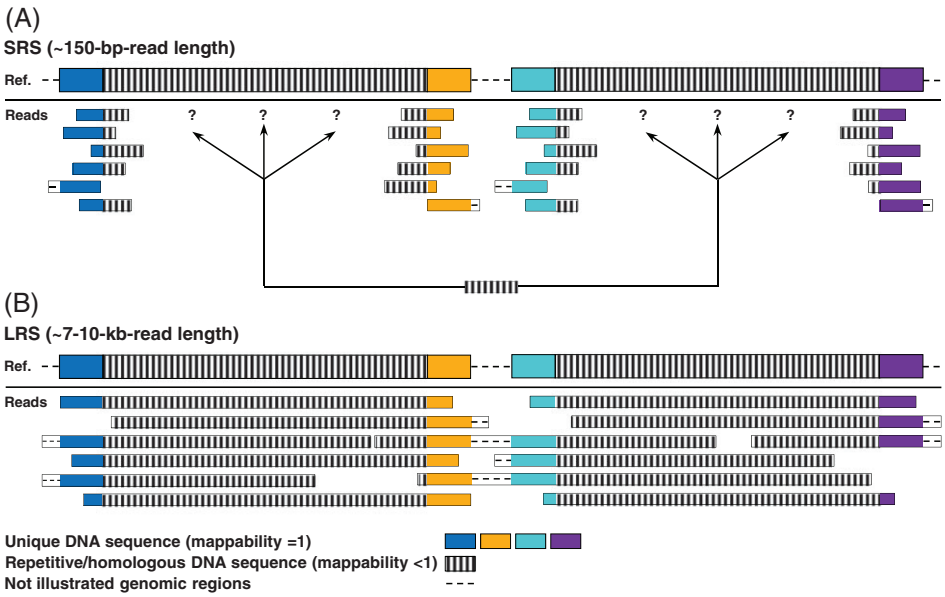

FIGURE 3 Schematic representation of read alignment when using SRS compared to LRS. (A) SRS is exemplified by 150-bp short reads. Note that repetitive/homologous regions longer than the read length cause ambiguous alignment, that is, mappability <1. (B) LRS is exemplified by multi-kb long reads. Note that long reads can cover unique DNA sequences flanking repetitive/homologous regions, enabling unambiguous alignment, that is, mappability = 1. LRS, long-read sequencing; Ref., reference genome; SRS, short-read sequencing

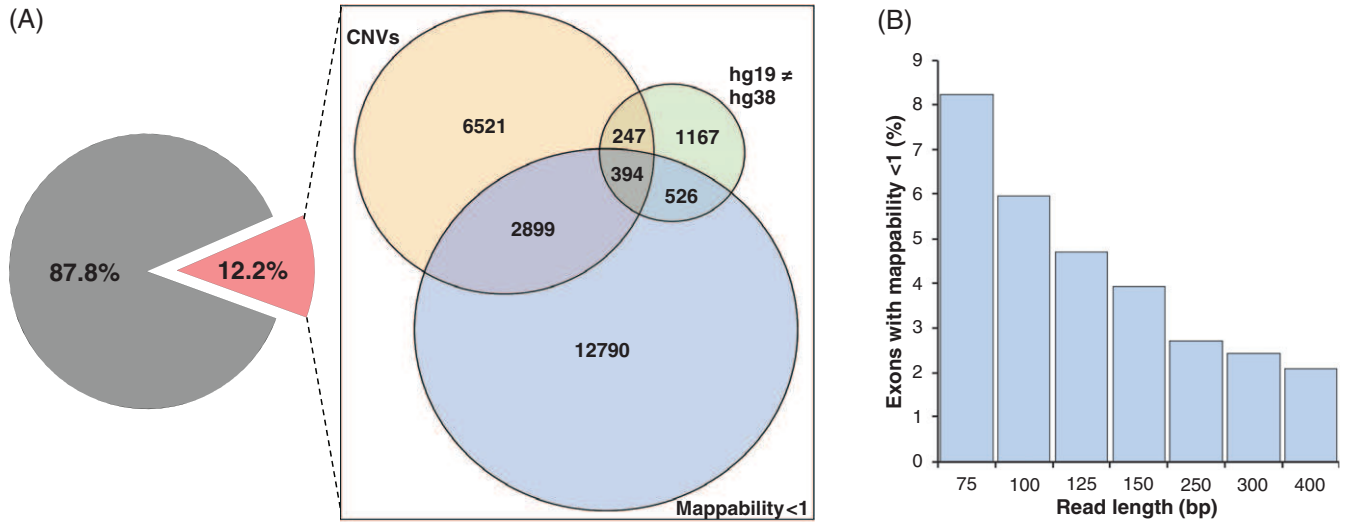

FIGURE 4 Short-read sequencing of RefSeq coding exons. (A) Percentage or number of exons with potentially affected (red) and unaffected (gray) read depth alignment and/or variant calling in short-read, whole-genome sequencing due to ambiguous 75-mer mappability (mappability<1), the presence of common copy number variations (CNVs) and the difference between the GRCh37 (ncbi.nlm.nih.gov/assembly/GCF_000001405.13) and GRCh38 (ncbi.nlm.nih.gov/assembly/GCF_000001405.26) reference genomes (hg19 ≠ hg38). (B) Percentage of exons with mappability<1 (calculated using GEM[33] version GEM-binaries-Linux-x86_64-20100419-003425 with m = 2 like the UCSC mappability tracks). Y-chromosomal exons were excluded. CNV, copy number variation

all SRS platforms is the short read length, leading to read alignment difficulties (Figure 3). Indeed, bioinformatic alignment of reads matching to more than 1 genomic location due to sequence homology (mappability <1) is ambiguous and can lead to misalignments as well as to false-positive and false-negative variant calling.[33–35]

To increase the awareness of SRS limitations, we filtered all 201 461 unique autosomal and X-chromosomal coding exons listed in RefSeq (ncbi.nlm.nih.gov/refseq, version December 2013) by 3 different criteria potentially affecting SRS read depth and/or variant calling. The first criterion was mappability, identifying coding exons that have at least 1 base with mappability <1 in 75-bp short sequences. The second criterion was the presence of common CNVs overlapping with the analyzed autosomal and X-chromosomal coding exons because it may lead to a lower (deletions) or higher (duplications) read depth, which may affect the calling of non-CNV variants. We therefore filtered out exons overlapping with common CNVs listed in the Database of Genomic Variants (DGV, dgv.tcag.ca, version July 2015)[36] detected in >90 individuals. The third criterion identified coding exons with errors in the reference genome, that is, updated in GRCh38/hg38 compared to GRCh37/hg19, as these can lead to erroneous alignment and variant calling. According to all 3 criteria, in ~12% of all coding exons listed in RefSeq special caution is needed in variant calling and interpretation of SRS data (Figure 4). Moreover, ~2% of all exons have 100% of positions or at least 1 250-bp-long contiguous region with mappability <1 defining the SRS dead zone[35] as well as ~17% of all 75-bp-long genomic regions have a mappability <1. Taken together, in addition to SRS appropriate complementary approaches such as long-read sequencing (LRS) have to be considered, at least for regions with mappability <1.

2.5 | Long-read sequencing

The read and mappability limitations of SRS can be overcome by using LRS. LRS with multi-kb-long reads facilitates unambiguous alignment to a reference genome due to the increased ability of long reads to completely span large complex, repetitive or homologous regions (Figure 3). By using LRS, the unambiguously mappable region of the genome can therefore be increased, minimizing clinically relevant false-negative results.[37] So far, 2 LRS technologies have been introduced, with real and synthetic long reads.

Real LRS, also known as third-generation sequencing, is typified by single-molecule sequencing of Pacific Biosciences (PacBio, pacb.com) as a market leader and Oxford Nanopore Technologies (ONT, nanoporetech.com) as a newcomer, both of which are able to sequence GC-rich regions with uniform coverage and to determine DNA base modifications without special treatment. PacBio uses single-molecule, real-time sequencing, generating reads with an average length of 10 to 15 kb.[38] PacBio sequencers have been used for different purposes such as de novo genome assembly, accurate genotyping of difficult-to-sequence regions, detection of CNVs and other SVs, discovery of transcriptome complexity and novel isoforms, and detection of methylation status.[37–40] The latest PacBio sequencer (Sequel) has been introduced to increase throughput and cost effectiveness but both remain considerably lower than in SRS.

The PacBio competitor ONT exploits the voltage change difference among nucleotides passing through a biological nanopore. ONT reads are on average 7 to 8 kb long[41] and can be generated on devices of different sizes such as the portable mobile-phone-sized MinION (1 flowcell) or the benchtop PromethION (48 flowcells) capable of sequencing a whole human genome.[42] In addition to PacBio-like applications,[41,43,44] the MinION has successfully been used for infectious disease outbreak surveillance in remote regions[45,46] and for phasing of cancer-related gene amplicons with especially high read depth comparable to short-read TS.[47,48] Despite having clear advantages in mappability and epigenetic mark detection over SRS, real LRS is not yet routinely applied due to its significantly lower throughput and higher per sample sequencing costs.[49] Furthermore,

PacBio and ONT both have high raw read error rates of ~10%.[38,41] As such errors are distributed randomly, their effect on consensus sequence and variant calling, however, can be minimized by increasing read depth and reading a template molecule multiple times (PacBio) or twice (ONT).[49] On the horizon are other real LRS platforms such as the mobile-phone-attachable SmidgION from ONT, solid-state nanopores from Hitachi[50] and the former Genia platform now owned by Roche (sequencing.roche.com), promising new possibilities for clinical applications.

Synthetic LRS is typified by long-read library preparation of Illumina (TruSeq Synthetic Long-Read DNA Library Preparation Kit, illumina.com) or 10X Genomics (10xgenomics.com), indexing sheared long fragments prior to SRS and bioinformatically merging short to long reads. As synthetic LRS is available on an SRS platform for a fraction of the cost, while offering similar results as real long reads, synthetic LRS is a promising alternative to real LRS. For example, synthetic long reads have successfully been applied to genome haplotyping and diploid de novo assemblies,[51,52] discovery of large genomic inversions,[53] and genome-wide reconstruction of complex SVs.[54] Although synthetic LRS may miss regions detectable by real LRS, it is predictable that LRS, synthetic or real, will soon find broad applications in clinical genetics.

3 | READ ALIGNMENT AND VARIANT CALLING

To detect sequence variants, the HTS output (short or long reads), stored in FASTQ or related formats, requires unambiguous read alignment to the reference genome (generating SAM/BAM files) and accurate variant calling (stored in (g)VCF files) using appropriate bioinformatic pipelines. As bioinformatic tools are typically specialized in a certain step of data analysis or have different performance,[55,56] a combination of appropriate tools is recommended to detect different types of sequence variants and to minimize clinically relevant false-negative results (ie, to zero as missing the disease-causing sequence variant leads to missing the molecular diagnosis). While the performance of HTS platforms has frequently been addressed in the last years, the performance of bioinformatic tools, especially for variant interpretation and LRS data analysis, is still a matter of ongoing research involving benchmark reference materials.[57–61]

3.1 | SNVs and insertions/deletions

The current de facto standard for the alignment of short reads and the calling of SNVs and indels is the BWA/GATK best practices pipeline,[62] which combines the read aligner BWA-MEM[63] and the variant caller GATK-HC.[64] As the computation time (~93 hours) and disk footprint (~254 GB) of BWA/GATK (version 3.5) is less suitable for large WGS data sets (60×), alternative tools such as Illumina's Isaac[65] and GENALICE MAP (genalice.com) have been introduced, offering faster analyses with reduced output file sizes.[60] Indeed, in our benchmarking study, GENALICE MAP showed ultrarapid speed (~1 hour) and superior low disk footprint (~6 GB) with BWA/GATK-like sensitivity for 60× short-read WGS data.[60] However, the run time of BWA/GATK may be accelerated 5× by the upcoming GATK version 4.0 and/or the DRAGEN platform (edicogenome.com) or compressive methods such as CORA,[66] allowing current BWA/GATK users to move on from TS or WES to WGS data analyses.

3.2 | Copy number variations

CNVs and copy neutral rearrangements such as inversions and translocations are SVs affecting large genomic segments.[67] Except for the software GROM (Genome Rearrangement OmniMapper), a recently introduced all-in-one solution to detect SNVs, indels, CNVs, and other SVs,[68] algorithms used for calling of SNVs and indels are not suited for the detection of larger sequence variants, requiring dedicated algorithms for CNV detection. Calling of CNVs from HTS data can be achieved by different approaches such as read-depth analysis, paired-end mapping, split reads, and de novo assembly.[24] A wide variety of CNV-calling tools uses 1 or several of these strategies, for example, CNVnator (read depth),[69] BreakDancer (paired-end mapping),[70] LUMPY and Manta (paired-end mapping and split reads),[71,72] or Cortex (de novo assembly).[73] Except for read-depth-based tools, CNV calling strategies are only applicable for WGS data, enabling the detection not only of CNVs but also of copy neutral SVs at base-pair resolution.[74] For CNV detection in TS and WES data, dedicated tools have been developed,[75] taking the enrichment bias into account, however, without achieving the accuracy observed in WGS data.[24,76] As CNV calling tools differ in their performance and hence users need to combine multiple algorithms, no de facto standard has been established for CNV detection.

3.3 | Short tandem repeats

STR expansions are highly relevant in clinical genetics but difficult to genotype accurately by SRS due to low-quality calls, very high GC content (as in fragile X syndrome) and/or expansions longer than the read length leading to ambiguous alignments that can be overcome by LRS. A number of software tools attempting to overcome SRS challenges for STR genotyping have recently been introduced, such as lobSTR,[77] ExpansionHunter,[78] HipSTR,[79] and STRetch,[80] demonstrating that also SRS data can lead to useful STR profiles. While the interpretation of STR expansions with known disease association is straightforward, the variant filtering and interpretation of other sequence variant types can be challenging.

4 | VARIANT FILTERING AND INTERPRETATION

HTS outputs about one called variant per 1000 bp of sequenced genome compared to the reference genome, leading to ten thousands and millions of sequence variants in WES and WGS, respectively. For time-consuming manual evaluation, appropriate filtering can help to reduce these large numbers by distinguishing pathogenic sequence variants from (likely) benign ones (Figure 1).[81] In silico gene panels can restrict sequence variants to the genes of interest. Gene-disease and variant-disease associations may be found in the literature (ncbi.

nlm.nih.gov/pubmed) and databases such as OMIM (omim.org), HGMD (portal.biobase-international.com), ClinVar (ncbi.nlm.nih.gov/clinvar), and LSDB (hgvs.org/locus-specific-mutation-databases). SNVs and indels can be prioritized based on their population frequency (eg, ExAC/gnomAD, gnomad.broadinstitute.org), phylogenetic conservation (eg, PhastCons, SiPhy), and potential effect on splicing (eg, Ada and RF scores in dbscSNV),[82] protein function or structure (eg, SIFT, PolyPhen2, MutationTaster2). For the prioritization of CNVs, their size and population frequency (eg, DGV) can be used. However, prioritization scores and predicted variant effects should never be automatically associated with pathogenicity, requiring manual expert review and intepretation.[83]

Several filtering and interpretation tools exist for the prioritization of annotated SNV and indel variants,[84] many of which have been reviewed elsewhere.[81,83] In contrast, only a few dedicated software tools are available for ranking the large amount of CNVs detected by genome-wide HTS (s. interpretation software examples in Table 1).[85] Intergenic CNVs disrupting the boundaries of topologically associating domains (TADs), in which genes have a higher probability of physical interaction,[86] should be included into a comprehensive CNV interpretation. In the interpretation of pathogenicity (especially in segregation analyses and genetic counseling), disease onset, parental mosaicism, dual molecular diagnoses (modifier), and random autosomal monoallelic expression[87,88] (Table 1) should also be considered. Moreover, although the accurate prediction of splicing defect is currently challenging and needs the development of better in silico tools in the future, it is clinically relevant to consider the possibility of aberrant splicing for any intronic and exonic sequence variants detected in diseases.[89–94] As a considerable number of new gene-disease and variant-disease associations have been reported annually, the regular reevaluation of HTS data using current knowledge is indicated for unsolved cases with suspected Mendelian disorder.[10,11]

Trio and segregation analyses, if available, can be carried out not only to provide further evidence for the disease association of detected sequence variant(s) but also to determine whether sequence variants are de novo or inherited, thereby implicating and enabling the recognition of at-risk relatives. Furthermore, the interpretation of rare or novel sequence variants might be facilitated by collaborative efforts and global-scale data sharing of geno- and phenotypic profiles, which can bring together identical or similar cases and hence increase significance through recurrent findings. One such possible platform is Matchmaker Exchange (matchmakerexchange.org), which includes multiple appropriate databases (eg, DECIPHER, decipher.sanger.ac.uk) and matching tools (eg, Gene Matcher, genematcher.org), allowing the connection between patients, clinicians, and researchers from around the world.[93] Successful data sharing, however, requires the use of standardized terms for describing clinical phenotypes, for which initiatives have been developed such as the Human Phenotype Ontology (HPO, human-phenotype-ontology.github.io).

The most comprehensive interpretation of sequence variants requests not only well-curated databases, segregation analyses, and data sharing but also appropriate functional analyses, for example, to assess the effect of variants on splicing and/or gene expression.[29,94,95] Nevertheless, despite the most comprehensive interpretation, variants of unknown significance (VUS) will remain one of the most frequent entities for years in the current genomics era.[81] Sequence variants to be reported might be confirmed by Sanger sequencing (for SNVs, indels, and breakpoints of large deletions), MLPA or quantitative PCR (for CNVs) and/or fragment length analyses (for STRs). Readers of genetic testing reports should bear in mind that sequencing technologies, variant calling, filtering, and interpretation differ among laboratories[94] as well as that incorrect connection between sequence variant and disease could have fatal consequences to patient health and family planning.[83]

5 | REGULATORY ISSUES: GENETIC COUNSELING, REIMBURSEMENT AND DATA STORAGE SAFETY

For clinical genetic testing with the purpose to identify the disease-causing sequence variant(s), an informed consent of the person being tested (or his/her legal representative) is required. Clinical genetic testing (not to be confused with direct-to-consumer genetic testing) should be preceded and followed by professional genetic counseling (acpm.org/?GeneticTestgClinRef) including appropriate information on incidental findings unrelated to the clinical indication but of medical value or utility, for example, for the prevention of a late-onset disease or for the identification and counseling of at-risk relatives. The reporting of incidental findings in WES/WGS is a challenging subject, for which the patient's decision should be respected and the guidelines established by the ACMG may be consulted.[18,19]

The reimbursement for clinical genetic testing varies among countries. Even when ordered by a qualified person (eg, medical geneticist, primary care doctor) it is recommended to inquire of the insurance company beforehand if costs are covered (ghr.nlm.nih.gov/primer/testing/insurancecoverage). Furthermore, when it comes to insurances there is a risk for genetic discrimination, which needs to be prevented.[96] Regulations are required in a way that both, patients and the healthcare systems, can benefit from the scientific advances.[97] The goal should be toward the benefit of patients in need while still being economically feasible.

With ever-increasing amounts of genetic data there are emerging regulatory issues regarding data safety and storage. Although cloud environments enable high-capacity data storage (eg, Amazon Web Services, aws.amazon.com), there are legitimate concerns about the privacy and security of sensitive cloud-stored genetic data,[98] indicating the need for more secure and decentralized solutions (eg, UnLynx).[99] Data sharing, which facilitates genotype-phenotype correlation, is subject to further discussions as it has been shown that even after de-identification, re-identification of a single person is at least partially possible by either STR[100] or SNP genotyping.[101]

6 | CONCLUSIONS AND PERSPECTIVES

The workflow used in HTS impacts the diagnostic yield. Contrary to expectations, whole does not equal whole in short-read WES and

WGS. From a technical point of view, PCR-free WGS is not only the most comprehensive SRS method but also the better WES as it provides the most uniform/complete coverage of the genome and covers the exome better than WES.[7] As a considerable number of coding exons remain insufficiently covered by SRS, LRS is envisioned as the future of sequencing. Until then, however, LRS could complement SRS by sequencing of insufficiently covered regions, for the selection of which CRISPR-Cas9-targeted enrichment could be considered.[105] In addition to LRS, next-generation genome mapping on nanochannel arrays (bionanogenomics.com) may enter clinical genetics enabling the detection of large SVs.[102] In the next years, the detection of CNVs and STR expansions by appropriate HTS methods as well as the combination of DNA and transcriptome sequencing will increase the diagnostic yield.[103]

Fueled by technological advances, HTS has led to the current situation in which our ability to sequence is greater than our ability to interpret the detected sequence variants. Indeed, the effects of VUS and non-exonic sequence variants pose a challenge for variant interpretation. Methods used for alignment, variant calling, and filtering can considerably influence sequence variant detection. Thus, it is recommended to use more than 1 independent data analysis pipeline and to reevaluate unsolved cases to minimize clinically relevant false-negative results. Exclusively relying on in silico interpretation tools is dangerous, because, although powerful, they are not without fault and not a replacement for expert review and functional analyses.[104] HTS for clinical applications has not yet reached its full potential, because of remaining technical (hardware and software) and regulatory issues. Especially an appropriate solution will have to be found for data storage and safety because of ever-increasing amounts of genomic data with lifetime value.

Conflict of interest

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported by the Gottfried & Julia Bangerter-Rhyner-Stiftung, COFRA Foundation, Ernst Göhner Stiftung, Foundation Suyana, Gebauer Stiftung, NOMIS Foundation, Palatin-Stiftung, and Schäppi-Jecklin Stiftung.

ORCID

G. Matyas http://orcid.org/0000-0002-3212-9963

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*. 1977;74:5463-5467.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17:333-351.
3. Feliubadaló L, Tonda R, Gausachs M, et al. Benchmarking of whole exome sequencing and ad hoc designed panels for genetic testing of hereditary cancer. *Sci Rep*. 2017;7:37984.
4. García-García G, Baux D, Faugère V, et al. Assessment of the latest NGS enrichment capture methods in clinical context. *Sci Rep*. 2016;6:20948.
5. Hehir-Kwa JY, Claustres M, Hastings RJ, et al. Towards a European consensus for reporting incidental findings during clinical NGS testing. *Eur J Hum Genet*. 2015;23:1601-1606.
6. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al. ACMG clinical laboratory standards for next-generation sequencing. *Genet Med*. 2013;15:733-747.
7. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet*. 2016;135:359-362.
8. Clamp M, Fry B, Kamal M, et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A*. 2007;104:19428-19433.
9. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*. 2009;106:19096-19101.
10. Eldomery MK, Coban-Akdemir Z, Harel T, et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med*. 2017;9:26.
11. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med*. 2017;19:209-214.
12. Alazami AM, Patel N, Shamseldin HE, et al. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep*. 2015;10:148-161.
13. Zhu X, Petrovski S, Xie P, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med*. 2015;17:774-781.
14. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur J Hum Genet*. 2017;25:176-182.
15. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genet Med*. 2016;18:696-704.
16. Lee H, Deignan JL, Dorrani N, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*. 2014;312:1880-1887.
17. Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. *Genet Med*. 2015;17:578-586.
18. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*. 2013;15:565-574.
19. Kalia SS, Adelman K, Bale SJ, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017;19:249-255.
20. Meienberg J, Zerjavic K, Keller I, et al. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res*. 2015;43:e76.
21. Belkadi A, Bolze A, Itan Y, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*. 2015;112:5473-5478.
22. Lelieveld SH, Spielmann M, Mundlos S, Veltman JA, Gilissen C. Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum Mutat*. 2015;36:815-822.
23. Farooqi MS, Mitui M, Londin ER, Park JY. High concentration capture probes enhance massively parallel sequencing assays. *Clin Chem*. 2016;62:1032-1034.
24. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol*. 2015;3:92.
25. Soden SE, Saunders CJ, Willig LK, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci Transl Med*. 2014;6:265ra168.
26. Yuen RK, Thiruvahindrapuram B, Merico D, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat Med*. 2015;21:185-191.

27. Stavropoulos DJ, Merico D, Jobling R, et al. Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. *NPJ Genom Med*. 2016;1:15012.
28. Lionel AC, Costain G, Monfared N, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med*. 2017; (Epub ahead of print). https://doi:10.1038/gim.2017.119.
29. Kremer LS, Bader DM, Mertes C, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun*. 2017;8:15824.
30. Drew L. Pharmacogenetics: the right drug for you. *Nature*. 2016; 537:60-62.
31. Owen RP, Gong L, Sagreiya H, Klein TE, Altman RB. VKORC1 pharmacogenomics summary. *Pharmacogenet Genomics*. 2010;20: 642-644.
32. Rasmussen HB, Dahmcke CM. Genome-wide identification of structural variants in genes encoding drug targets: possible implications for individualized drug therapy. *Pharmacogenet Genomics*. 2012;22: 471-483.
33. Derrien T, Estellé J, Marco Sola S, et al. Fast computation and applications of genome mappability. *PLoS One*. 2012;7:e30377.
34. Goldfeder RL, Priest JR, Zook JM, et al. Medical implications of technical accuracy in genome sequencing. *Genome Med*. 2016;8:24.
35. Mandelker D, Schmidt RJ, Ankala A, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet Med*. 2016;18:1282-1289.
36. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42: D986-D992.
37. Merker JD, Wenger AM, Sneddon T, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med*. 2018;20:159-163.
38. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics*. 2015;13:278-289.
39. Weirather JL, Afshar PT, Clark TA, et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Res*. 2015;43:e116.
40. Seo JS, Rhie A, Kim J, et al. De novo assembly and phasing of a Korean human genome. *Nature*. 2016;538:243-247.
41. Leggett RM, Clark MD. A world of opportunities with nanopore sequencing. *J Exp Bot*. 2017;68:5419-5429.
42. Loose MW. The potential impact of nanopore sequencing on human genetics. *Hum Mol Genet*. 2017;26:202-207.
43. Jain M, Koren S, Quick J, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *bioRxiv*. 2017. https://doi.org/10.1101/128835.
44. Simpson JT, Workman RE, Zuzarte PC, Matei D, Durs LJ, Timp W. Detecting DNAcytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14:407-410.
45. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016;530:228-232.
46. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. 2017; 544:309-315.
47. Suzuki A, Suzuki M, Mizushima-Sugano J, et al. Sequencing and phasing cancer mutations in lung cancers using a long-read portable sequencer. *DNA Res*. 2017;24:585-596.
48. de Jong LC, Cree S, Lattimore V, et al. Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. *Breast Cancer Res*. 2017;19:127.
49. Weirather JL, de Cesare M, Wang Y, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore technologies and their applications to transcriptome analysis. *F1000Res*. 2017; 6:100.
50. Goto Y, Yanagi I, Matsui K, Yokoi T, Takeda K. Integrated solid-state nanopore platform for nanopore fabrication via dielectric breakdown, DNA-speed deceleration and noise reduction. *Sci Rep*. 2016; 6:31324.
51. Zheng GX, Lau BT, Schnall-Levin M, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol*. 2016;34:303-311.
52. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. *Genome Res*. 2017;27: 757-767.
53. Eslami Rasekh M, Chiatante G, Miroballo M, et al. Discovery of large genomic inversions using long range information. *BMC Genomics*. 2017;18:65.
54. Spies N, Weng Z, Bishara A, et al. Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods*. 2017; 14:915-920.
55. O'Rawe J, Jiang T, Sun G, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med*. 2013;5:28.
56. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant-calling pipelines using gold standard personal exome variants. *Sci Rep*. 2015;5:17875.
57. Liu X, Han S, Wang Z, Gelernter J, Yang BZ. Variant callers for next-generation sequencing data: a comparison study. *PLoS One*. 2013;8:e75619.
58. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*. 2014;15:256-278.
59. Salgado D, Bellgard MI, Desvignes JP, Béroud C. How to identify pathogenic mutations among all those variations: variant annotation and filtration in the genome sequencing era. *Human Mutat*. 2016;37:1272-1282.
60. Plüss M, Kopps AM, Keller I, et al. Need for speed in accurate whole-genome data analysis: GENALICE MAP challenges BWA/-GATK more than PEMapper/PECaller and Isaac. *Proc Natl Acad Sci U S A*. 2017;114:E8320-E8322.
61. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data*. 2016;3:160025.
62. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43: 11.10.1-11.1033.
63. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. 2013;1303:3997v2.
64. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297-1303.
65. Raczy C, Petrovski R, Saunders CT, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29:2041-2043.
66. Yorukoglu D, YW Y, Peng J, Berger B. Compressive mapping for next-generation sequencing. *Nat Biotechnol*. 2016;34:374-376.
67. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: new insights in genome diversity. *Genome Res*. 2006;16:949-961.
68. Smith SD, Kawash JK, Grigoriev A. Lightning-fast genome variant detection with GROM. *GigaScience*. 2017;6:1-7.
69. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011; 21:974-984.
70. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009;6:677-681.
71. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014; 15:R84.
72. Chen XY, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016;32:1220-1222.
73. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*. 2012;44:226-232.
74. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14:S1.
75. Zare F, Dow M, Monteleone N, Hosny A, Nabavi S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*. 2017;18:286.

76. Tan R, Wang Y, Kleinstein SE, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat*. 2014;35:899-907.
77. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res*. 2012;22:1154-1162.
78. Dolzhenko E, van Vugt JJFA, Shaw RJ, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res*. 2017;27:1895-1903.
79. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14:590-592.
80. Dashnow H, Lek M, Phipson B, et al. STRetch: detecting and discovering pathogenic short tandem repeats expansions. *bioRxiv*. 2017. https://doi.org/10.1101/159228.
81. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17:405-424.
82. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res*. 2014;42:13534-13544.
83. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet*. 2017;18:599-612.
84. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
85. Nowakowska B. Clinical interpretation of copy number variants in the human genome. *J Appl Genet*. 2017;58:449-457.
86. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*. 2012;485:376-380.
87. Ohlsson R. Genetics. Widespread monoallelic expression. *Science*. 2007;318:1077-1078.
88. Savova V, Patsenker J, Vigneau S, Gimelbrant AA. dbMAE: the database of autosomal monoallelic expression. *Nucleic Acids Res*. 2016;44:753-756.
89. Houdayer C, Caux-Moncoutier V, Krieger S, et al. Guidelines for splicing analysis in molecular diagnosis derived from a set of 327 combined in silico/in vitro studies on BRCA1 and BRCA2 variants. *Hum Mutat*. 2012;33:1228-1238.
90. Caminsky NG, Mucaki EJ, Rogan PK. Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Res*. 2014;3:282.
91. Ohno K, Takeda JI, Masuda A. Rules and tools to predict the splicing effects of exonic and intronic mutations. *Wiley Interdiscip Rev RNA*. 2017;e1451. https://doi.org/10.1002/wrna.1451.
92. Soens ZT, Branch J, Wu S, et al. Leveraging splice-affecting variant predictors and a minigene validation system to identify Mendelian disease-causing variants among exon-captured variants of uncertain significance. *Hum Mutat*. 2017;38:1521-1533.
93. Philippakis AA, Azzariti DR, Beltran S, et al. The matchmaker exchange: a platform for rare disease gene discovery. *Hum Mutat*. 2015;36:915-921.
94. Pepin MG, Murray ML, Bailey S, Leistritz-Kessle G, Schwarze U, Byers PH. The challenge of comprehensive and consistent sequence variant interpretation between clinical laboratories. *Genet Med*. 2016;18:20-24.
95. Woods NT, Baskin R, Golubeva V, et al. Functional assays provide a robust tool for the clinical annotation of genetic variants of uncertain significance. *NPJ Genom Med*. 2016;1:16001.
96. Erwin C. Legal update: living with the genetic information nondiscrimination act. *Genet Med*. 2008;10:869-873.
97. Vozikis A, Cooper DN, Mitropoulou C, et al. Test pricing and reimbursement in genomic medicine: towards a general strategy. *Pub Health Genomics*. 2016;19:352-363.
98. Tang H, Jiang X, Wang X, et al. Protecting genomic data analytics in the cloud: state of the art and opportunities. *BMC Med Genomics*. 2016;9:63.
99. Froelicher D, Egger P, Sá Sousa J, et al. UnLynx: a decentralized system for privacy-conscious data sharing. *Proc Privacy Enhancing Technol*. 2017;4:152-170.
100. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339:321-324.
101. Lippert C, Sabatini R, Maher MC, et al. Identification of individuals by trait prediction using whole-genome sequencing data. *Proc Natl Acad Sci U S A*. 2017;114:10166-10171.
102. Jaratlerdsiri W, Chan EKF, Petersen DC, et al. Next generation mapping reveals novel large genomic rearrangements in prostate cancer. *Oncotarget*. 2017;8:23588-23602.
103. Cummings BB, Marshall JL, Tukiainen T, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med*. 2017;9:386.
104. Miosge LA, Field MA, Sontani Y, et al. Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci U S A*. 2015;112:e5189-e5198.
105. Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat Commun*. 2017;8:14291.
106. Guo Y, Gifford DK. Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. *BMC Genomics*. 2017;18:45.
107. Klein, A. Hard Drive Cost Per Gigabyte. backblaze.com/blog/hard-drive-cost-per-gigabyte. Accessed November 23, 2017.

How to cite this article: Caspar SM, Dubacher N, Kopps AM, Meienberg J, Henggeler C, Matyas G. Clinical sequencing: From raw data to diagnosis with lifetime value. *Clin Genet*. 2018;93:508–519. https://doi.org/10.1111/cge.13190